Urban Re-Identification: Fusing Local and Global Features with Residual Masked Maps for Enhanced Vehicle Monitoring in Small Datasets

William A. Ramirez[a], Cesar A. Sierra Franco[b], Thiago R. da Motta[c] and Alberto Raposo[d]
Pontifical Catholic University of Rio de Janeiro, Brazil

Keywords: Urban Re-Identification, Dilated Region Proposal, Local and Global Attribute Fusion, Residual Connection Modules, Multi-Object Tracking.

Abstract: This paper presents an optimized vehicle re-identification (Re-ID) approach focused on small datasets. While most existing literature concentrates on deep learning techniques applied to large datasets, this work addresses the specific challenges of working with smaller datasets, mainly when dealing with incomplete partitioning information. Our approach explores automated regional proposal methods, examining residuality and uniform sampling techniques for connected regions through statistical methods. Additionally, we integrate global and local attributes based on mask extraction to improve the generalization of the learning process. This led to a more effective balance between small and large datasets, achieving up to an 8.3% improvement in Cumulative Matching Characteristics (CMC) at k=5 compared to attention-based methods for small datasets. We improved generalization regarding context changes of up to 13% in CMC for large datasets. The code, model, and DeepStream-based implementations are available at https://github.com/will9426/will9426-automatic-Region-proposal-for-cars-in-Re-id-models.

1 INTRODUCTION

Re-identification in urban areas, and generally in uncommon classes, often falls into small dataset scenarios. In these cases, video frames with a temporal relationship and tracking information from one or multiple cameras are commonly considered. This field can involve various types of approaches categorized into Mask-Guided Models (Song et al., 2018; Kalayeh et al., 2018; Zhang et al., 2020; Lv et al., 2024), Stripe-Based Methods (Luo et al., 2019b; Wang et al., 2018; Fan et al., 2019), Attention-Based Methods (Si et al., 2018; Chen et al., 2021), and GAN-Based Methods (Jiang et al., 2021), often supplemented by re-ranking methods (Luo et al., 2019a). The previously mentioned techniques result from challenges focused on data availability and the need to extract increasingly deeper and more representative attributes. Mask-guided models often entail additional costs in annotation or inference, as it is necessary to have a process for detecting parts of the instances. (Zhang et al., 2020) shows how this can involve an additional annotation task in exchange for optimizing the identification process due to the availability of more contextual information. On the other hand, Stripe-Based Methods highlight another possible approach to global and localized attribute extraction, using statistical methods to propose areas that may be more representative, shifting the annotation cost to a method embedded within the model. Regarding the use of attention modules along with partition-based techniques, (Zhang et al., 2020) points out the fusion of local and global attributes using squeeze and excitation layers (SElayers) and attention, achieving improvements of up to 1.2% in the Cumulative Matching Characteristics (CMC), a metric specialized in re-identification. Each of these methods presents trade-offs: Mask-guided models offer fine-grained detail but require extensive annotation, increasing complexity; Stripe-Based Methods simplify this at the risk of oversimplifying regions of interest; Attention-Based Methods balance attribute fusion but may overfit or add computational overhead; and GAN-Based Methods provide flexibility in augmenting data but can be computationally intensive and challenging to stabi-

[a] https://orcid.org/0000-0003-1060-1523
[b] https://orcid.org/0000-0002-5825-8798
[c] https://orcid.org/0000-0002-9579-5867
[d] https://orcid.org/0000-0001-7279-1823

Ramirez, W. A., Franco, C. A. S., R. da Motta, T. and Raposo, A.
Urban Re-Identification: Fusing Local and Global Features with Residual Masked Maps for Enhanced Vehicle Monitoring in Small Datasets.
DOI: 10.5220/0013176300003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 574-581
ISBN: 978-989-758-728-3; ISSN: 2184-4321

lize.

This paper addresses the challenges associated with small datasets by studying algorithms for efficient region estimation and improving feature map diversity to enhance model generalization. We explore using statistical methods and attribute extraction techniques to create connected areas, proposing more representative regions or stripes that lead to feature maps with greater representativeness. The proposed model consists of three branches: localized feature maps (using RPN), globalized features (leveraging the backbone and our custom attribute extraction block), and base features from the backbone. Our case study uses the VRIC and VeRi datasets, including cross-validation experiments with mixed training and testing configurations, specifically Vric− >VeRi and VeRi− >Vric. Our approach builds on the baseline established in (Zhang et al., 2020; Luo et al., 2019a), which emphasizes the use of camera-guided partitioned attention and attribute fusion while discussing the limitations associated with this module.

In summary, the main contributions of our work are as follows:

- We propose an optimized approach for region extraction, demonstrating how the use of statistical methods like Monte Carlo can be helpful for stripe extraction while simultaneously showing how using the same backbone to propose regions can be a successful path to achieving a more balanced model in terms of response to variance.
- We propose a validation method for the trained model, introducing the concept of cross-inference between two datasets with the same category about the instance but with apparent differences in context and resolution. We aim to demonstrate that the trained model can be used in other contexts.
- Based on the challenges discussed around re-identification, we proposed a computationally balanced and reliable model to establish a real-time baseline using DeepStream.

1.1 Data Augmentation

Deep learning approaches are often highlighted for their reliance on large datasets. Data augmentation has emerged as a valuable strategy in the context of re-identification with small datasets. Techniques such as grayscale conversion, random erasing, image orientation flips, and zooming generate synthetic data with some variability (Gong et al., 2021; Jiang et al., 2021). Similarly, Generative Networks (Zheng et al., 2019; He et al., 2023; Karras et al., 2020; Karras et al., 2021) have proven effective in increasing variability in small datasets with low variability, facilitating data augmentation in labeled Re-identification datasets.

In the context of GANs, DG-Net (Zheng et al., 2019) emphasizes its ability to generalize key features such as pose and clothing. The model incorporates a feature disentangling module and reconstruction loss, enhancing cross-domain generation. DG-GAN (He et al., 2023) aims to learn from defects or irregular regions. DG-GAN architecture includes two generators and four discriminators across two domains, enhancing synthetic data generation for pattern recognition in land surfaces, such as roads or open areas. The StyleGAN architecture has gained attention in recent years for its contributions to synthetic datasets (Karras et al., 2020; Karras et al., 2021). StyleGAN3 (Karras et al., 2021), represents a significant advancement in computer vision, albeit with high computational demands. StyleGAN3 employs components like the Mapping Network, Synthesis Network, and Weight Demodulation to provide fine control over image style while effectively addressing aliasing, resulting in high-quality images without noise associated with generative learning.

1.2 Baseline for ReID

In the past decade, ReID methods have evolved into several categories: Mask-Guided Models, Stripe-Based Methods, Pose-Guided Methods, Attention-Based Methods, and GAN-Based Methods, often with re-ranking techniques to address data limitations and identification challenges. Mask-guided models use instance-specific masks derived from segmentation or detection (Song et al., 2018; Kalayeh et al., 2018; Zhang et al., 2020; Lv et al., 2024), combining local and global attributes. Despite the additional processing costs, they enhance matching performance, achieving up to 93% for CMC@5 (Zhang et al., 2020; Lv et al., 2024). Stripe-based methods segment images to create local embeddings, as seen in (Luo et al., 2019b; Wang et al., 2018; Fan et al., 2019; Fawad et al., 2020). However, alignment issues challenge these methods, leading to approaches like AlignedReID++ (Luo et al., 2019b), which reported a 3% improvement in Rank-1 accuracy compared to models without alignment. Attention-based methods extract discriminative features without explicit masks (Si et al., 2018; Chen et al., 2021). While effective for large datasets, they face overfitting issues in smaller ones. For instance, (Chen et al., 2021) reported a 9% improvement in MAP@5 for larger datasets. Gan-based methods address small dataset limitations by generating synthetic data. For exam-

ple, (Jiang et al., 2021) enhanced person category diversity using GANs, achieving a 1% improvement in CMC@1. Batch Normalization improves training stability and generalization for large datasets. (Luo et al., 2019a) demonstrated up to a 6% enhancement for CMC@1.

2 METHOD

2.1 Dilated Region Proposal for Cars (DRPC)

We proposed a module inspired by (Chen et al., 2023; Lv et al., 2024), where we use the ResNet-50 backbone to propose regions. Initially, we use the first layer of our backbone, where we extract low-level features. In this layer, with tensor dimensions [B, C, H, W], a feature map is generated as $F = \text{ResNet}(input\_tensor)$. We then average across the channel dimension to obtain a 2D representation of the feature magnitudes, as shown in Eq. 1.

$$M_{avg}(b,h,w) = \frac{1}{C_{out}} \sum_{c=1}^{C_{out}} F(b,c,h,w) \quad (1)$$

Where M_{avg} is the average tensor with dimensions $[B,1,H_{out},W_{out}]$.

To create a candidate region based on Eq. 2, we use an adaptative threshold over the global average value of M_{avg}:

$$M_{avg}(b,h,w) = \begin{cases} 1 & \text{if } M_{avg}(b,h,w) > \mu_{avg} \\ 0 & \text{if } M_{avg}(b,h,w) \leq \mu_{avg} \end{cases} \quad (2)$$

Where μ_{avg} is the global average value of M_{avg} described on Eq. 3.

$$\mu_{avg} = \frac{1}{B \cdot H_{out} \cdot W_{out}} \sum_{b=1}^{B} \sum_{h=1}^{H_{out}} \sum_{w=1}^{W_{out}} M_{avg}(b,h,w) \quad (3)$$

Thus, we obtain a binary candidate mask that highlights areas of interest based on the feature map extracted. However, our goal is to create regions around these initially highlighted characteristics in stripes, aiming for the model to emphasize contours and other low-level features of the pre-selected area.

To create the areas, we initialize a mask $M_{regions}$ with dimensions $[B,N,H_{out},W_{out}]$, where N is the number of desired regions. Each vertical stripe is extracted and dilated. Consider a stripe R with dimensions $[H_{out},\text{width}]$, where 'width' is the width of the stripe. The dilation is performed using a structural operation that expands the stripe. If R is the original stripe and D is the result after dilating R, as shown in Eq. 4:

$$D_{dilated} = R \oplus K \quad (4)$$

Where $\oplus$ denotes the morphological dilation and K is a dilation kernel (in this case, a matrix of ones). Finally, the dilated stripe $D_{dilated}$ is assigned to the region mask $M_{regions}$ in the corresponding position, Eq. [5,6] describe the proposed region and the result in the extraction within the proposed area.

$$M_{regions}(b,i,:,\text{start}_i : \text{end}_i) = D_{dilated} \quad (5)$$

$$\mathbf{F}_n^{\text{prop}} = \mathbf{F} \odot M_{regions} \quad (6)$$

2.2 Quasi-Monte Carlo for Proposal Regions (QMCPR)

This section introduces the use of **Quasi-Monte Carlo (QMC)** to generate proposal regions without relying on pre-annotated masks or outputs from segmenters or detectors. Strips $\mathbf{x}_n$ are defined in a unit space $[0,1]^d$ using the Halton sequence, which distributes points uniformly. For the two-dimensional case $d = 2$, each strip $\mathbf{x}_n = (x_n, y_n)$ is derived using bases b_1 and b_2, as shown in Eq. 7.

$$\mathbf{x}_n = (\phi_{b_1}(n), \phi_{b_2}(n)) \quad (7)$$

where $\phi_b(n)$ is the inverse radical function in base b, which converts an integer n into a fraction in base b. Initially, the most statistically discriminative areas are selected through a histogram analysis of the instance-level image to restrict the search areas and highlight the discriminative areas. Based on this analysis, greater weight is given to the bases in the most relevant regions.

Subsequently, the QMC method is applied to distribute the strips in these areas, ensuring that the strips cover at least 15% of the total width or length of the image, which is crucial to guarantee that the strips contain sufficient visual context. This strategy is beneficial for providing variability of context or information to the model, allowing the model to relate the instance from a smaller context. The generated strips $\mathbf{x}_n$ are originally fractions within the range $[0,1]$. To adapt them to the image space of size $H \times W$ (where H and W are the dimensions of the image), these points are scaled to the appropriate range, as shown in Eq. 8.

$$\mathbf{p}_n = \mathbf{x}_n \times \begin{pmatrix} W - w_m \\ H - h_m \end{pmatrix} \quad (8)$$

where w_m and h_m are the dimensions of the proposed mask. The scaled values $\mathbf{p}_n = (p_x, p_y)$ are rounded to obtain the pixel indices $(x_{\text{start}}, y_{\text{start}})$ where the mask will begin in the image. For each generated strip $\mathbf{p}_n$, a binary strip mask M_n of size $w_m \times h_m$ is defined in an image of size $H \times W$. Eq. 9 shows the cases for generating the strip.

$$M_n(i,j) = \begin{cases} 1 & \text{if } \begin{matrix} x_{\text{start}} \leq i < x_{\text{start}} + w_m \\ y_{\text{start}} \leq j < y_{\text{start}} + h_m \end{matrix} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

Where (i, j) are the pixel indices in the image, the masks M_n are used to propose specific regions of the image that are then processed to extract a series of features. $\mathbf{F}$ is an image feature obtained from an intermediate layer of the network, so by applying the mask M_n on $\mathbf{F}$, a proposed region feature is generated. Eq. 10 describes how we use the proposed region to define our proposed feature map.

$$\mathbf{F}_n^{\text{prop}} = \mathbf{F} \odot M_n \quad (10)$$

where $\odot$ denotes the element-wise product. This proposed feature $\mathbf{F}_n^{\text{prop}}$ can be used to train the model by combining it with the global feature of the image, allowing the model to focus on both local and global features.

2.3 Feature Maps

The proposed model will address three branches of the extraction and fusion of attributes: a base feature map, a global feature map, and a local feature map. The base map will describe the most superficial attributes found in the initial layers of the network, obtained through convolutions at different levels. In this stage, our model aims to capture patterns that may be associated with edges, textures, and colors (low dimensionality). Eq. 11 shows what this branch represents.

$$\mathbf{f}_{\text{base}} = \sigma(\mathbf{W} * \mathbf{X} + \mathbf{b}) \quad (11)$$

where

$\mathbf{W}$ is the convolution kernel, $*$ represents the convolution operation, $\mathbf{X}$ is the input (image or previous features), $\mathbf{b}$ is the bias, and σ is an activation function (such as ReLU).

Concerning the global attributes, these are obtained by aggregating all the spatial information of the image into a single representative vector using *Global Average Pooling (GAP)*, just as shown in Eq. 12.

$$\mathbf{f}_{\text{global}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{f}_{ij} \quad (12)$$

where $\mathbf{f}_{ij}$ is the feature vector at the spatial position (i, j), and $H \times W$ is the size of the spatial feature.

In our small dataset context, f_{global} will contain representations that help prevent overfitting by not relying on specific details and providing a robust representation that tolerates minor variations in the image.

f_{proposal} will represent the local attributes extracted from the regions or stripes. In our case, M_{regions} will be the proposed stripes using the methods mentioned above, which will be used as delimiters to extract more localized attributes, $\mathbf{f}_{\text{proposal}}$ in Eq. 13.

$$\mathbf{f}_{\text{proposal}} = \sum_{M=1}^{M} \alpha_M \cdot \mathbf{f}_{(\text{Input},M)} \quad (13)$$

where α_M are the selection weights, and $\mathbf{f}_{\text{Input},M}$ are the features in M region . $\mathbf{f}_{\text{proposal}}$ generates representative regions obtained through the outputs of the DRPRC or QMCPR modules.

Subsequently, the reduced features are obtained through a reduction of dimensionality for the feature map of the proposed regions, performing an aggregation along with the activation, see Eq. 14.

$$\mathbf{f}_{\text{reduce}} = \sigma\left(\frac{1}{F} \sum_{f=1}^{F} \text{proposal}_{b,f}\right) \quad (14)$$

In this case, we calculate the mean of the values across the feature dimension for each batch b.

2.4 Set up

Regarding our loss metric, we considered a fusion of losses in triplet loss, starting with a cross-entropy loss related to the class, followed by a loss for base attributes, another for the proposed region, and another for global attributes. Figure 1 shows the meaning of each of these losses.

The total loss function (L_{total}) is given by the Eq. 15 where λ_{triplet} is defined in 16.

$$\begin{aligned} L_{\text{total}} = & \lambda_{\text{id}} \cdot L_{\text{CE}}(\text{cls}_i, y) \\ & + \lambda_{\text{triplet}} \cdot L_{\text{triplet}}(\text{global}_i, y) \\ & + pr_1 \cdot L_{\text{triplet}}(\text{prop}_i, y) \\ & + pr_2 \cdot L_{\text{triplet}}(\text{base}_i, y) \end{aligned} \quad (15)$$

$$L_{\text{triplet}} = \max(0, \text{dist}_{\text{ap}} - \text{dist}_{\text{an}} + \text{margin}) \quad (16)$$

Where:

λ_{id} is the weight assigned to the identification loss. λ_{triplet} is the weight assigned to the triplet loss. pr_1 is the weight assigned to the proposal loss. pr_2 is the

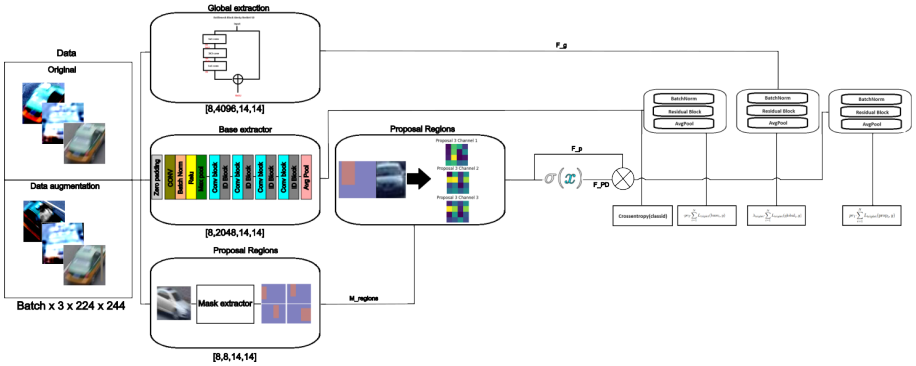

Figure 1: Reid training Pipeline.

weight assigned to the base loss. L_{CE} is the cross-entropy loss function, which measures the discrepancy between the class prediction and the target label. L_{triplet} is the triplet loss function, which measures the relative distance between an anchor, a positive, and a negative in the feature space. N is the number of layers or extracted features to which the loss is applied. cls_i refers to the classification outputs of each layer. global_i refers to the global features extracted from each layer. prop_i refers to the proposed features extracted from each layer. base_i refers to the base features extracted from each layer. y is the target label for classification. $dist_{ap}$ defines Anchor-Positive Distance and $dist_{an}$ defines Anchor-Negative Distance.

About our approach, we used the training configurations described in Table 1.

Table 1: Training Configuration.

Configuration	Value
Input Size	3x224x224
Epochs	200
Early Stop	CMC rank 1 tolerance (10 epochs)
Augmentation	Flip, Rotation, Scaling, Grayscale
Learning Rate	0.001
Optimizer	Adam
Reduce LR	On Plateau

3 EXPERIMENTS AND RESULTS

3.1 Datasets and Metrics

We use the VRIC dataset (Kanaci et al., 2018) in both small and large variants. The small version has 5,854 training samples (220 IDs, 10-30 samples/ID) and 2,811 gallery samples (validation), maintaining the original test set for consistent comparative analysis. The large version retains all 60,430 samples (5,622 IDs). Additionally, we use the VeRi dataset (Liu et al., 2016) to evaluate the model's adaptability to domain shifts. Our validation protocol considered cross-dataset evaluation of metrics to assess performance in two different scenarios: one with richer contextual information and the other with higher resolution.

The small dataset is designed to evaluate model performance with fewer, more variable inputs, testing generalization on unseen validation IDs. Data augmentation includes flips, grayscale conversions, area-specific and full rotations, and GAN-based view changes.

Performance is measured using mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) metrics, commonly used in re-identification tasks. mAP assesses retrieval effectiveness (Eq. 17), while CMC evaluates ranking accuracy (Eq. 18) for top $k = 1, 5, 10$.

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{k=1}^{m_i}P(k) \tag{17}$$

$$\text{CMC}(r) = \frac{1}{N}\sum_{i=1}^{N}\delta(\text{rank}_i \leq r) \tag{18}$$

3.2 Analysis of Attribute Extraction Method

The present work evaluated two methods for region generation. The first, DRPC, is based on generating dilated regions using contours from the first layer of the ResNet50 backbone, which involves a computational cost of $O(batch) \times processing1 + O(regions)$. In the first stage, we iterate over the entire batch and return the first layer in the forward pass, performing a pass through the model and returning the first convolution with binarization. The second stage encloses each binarized area based on a dilation. The second method, QMCPR, has a sublinear computational

cost and generates masks with an approximate computational cost of $O(batch \times regions)$, approximately $O(8n)$, where "regions" is a parameter fixed at 8, and "batch" is a variable input parameter. Both methods did not drastically modify the computational cost during training execution.

On the other hand, these two algorithms provide a base approach using masks that can be generalized to any ReID dataset, reducing the need for manual annotations or training models for part segmentation. The previously mentioned methods will enable the local extraction of features. Since these do not rely on pre-annotated masks, they will evaluate more representative regions using the ResNet backbone's upper layers or by selecting regions based on distributions within a unit field representing the area of interest. Subsequently, using extraction blocks, the features within these regions will be extracted and reduced to more relevant values using activations suggested in this work. The proposed use of regions enhances feature diversity by leveraging incomplete partitioning, which fuses and refines attributes from local, global, and base extractions. This approach enables the model to learn more discriminative contextual and instance-specific information in small datasets, making it a more robust solution.

Table 2 provides an initial notion of what these methods imply in small datasets. Here, we compare the model with mask guidance, the GRFR, and the DRPC model against the baseline PGAN model without attention. We observe that the proposed methods are quite close to the training done entirely with the masks generated for the VRIC dataset, provided in (Zhang et al., 2020), showing a correlation with the extraction of local attributes, which is a positive indicator that our method successfully proposes highly discriminative regions.

Table 2: Results of Different Models on Small-dataset Re-Identification Tasks.

Model	mAP	CMC Top 1	CMC Top 5
PGAN	45.8	33.7	60.1
model mask guided	53.9	43.2	66.2
model GRFR (ours)	53.5	43.1	67.3
model DRPC (ours)	56.0	44.5	68.3
model DRPC+AUG view (ours)	**63.3**	**53.3**	**74.9**

In the experimental stage, we observed that paying attention to small datasets tends to degrade the gradient, resulting in less model generalization. Comparing PGAN with and without attention (PGAN vs. model mask guided), we see that in small datasets, more globalized learning tends to bring better performance in terms of the CMC Top K=5 metric (approximately 6% improvement). In small datasets, the limited diversity and number of examples lead the attention mechanism to overfit specific, less generalizable local features rather than capturing broader patterns. This overfitting causes the model to become sensitive to noise or small variations in the training data, which in turn degrades the gradient during optimization and hinders the model's ability to generalize effectively to unseen samples. Regarding other masking methods, our trend was quite in line with PGAN without attention, in some cases having a better CMC K=5 (approximately 1.1% improvement).

3.3 Performance of Our Model

Our model is designed to handle small datasets without region annotations, using a model masking approach based on the forward pass of the same model. As depicted in Figure 1, we emphasize more representative areas of the feature maps through a reduction facilitated by an activation function. This approach reduced training time by several seconds due to its lower complexity, eliminating the need for attention mechanisms and specialized refinement modules. Specifically, the training times were 66 seconds per batch (size of 8 samples) for the base model, compared to 62 seconds with the DRPC method and 60 seconds with the QMCPR method. These tests were performed using an NVIDIA GeForce RTX 3060, a 13th Gen Intel® Core™ i5-13600KF (20 cores), and 32GB of RAM.

Table 3: Experimental results of Different Models on Large Dataset Re-Identification Tasks, $VRIC- > VRIC$.

Model	mAP	CMC Top 1	CMC Top 5
PGAN	**84.5**	**77.4**	**93.1**
model DRPC (ours)	82.0	74.9	89.8

As shown in Table 3, our model lags behind attention-based models on large datasets. This is because attention-based models typically involve deeper feature extraction layers. In contrast, models based on Squeeze-and-Excitation layers and attention mechanisms achieved up to a 3.3% improvement.

However, as illustrated in Table 4, our model demonstrated balanced performance across various scenarios due to our generalized approach in extracting local discriminative features and global attributes. Our model had fewer parameters, reducing inference times: 0.0302 seconds for the base model, 0.0095 seconds for QMPRC, and 0.0216 seconds for DPRC.

Table 4 details the performance of our model in mixed scenarios. The first scenario involves training and evaluating the VeRi model on the VeRi test set. This setup shows that VeRi, a higher-quality dataset, allows more straightforward methods that improve variability by expanding the visual context to

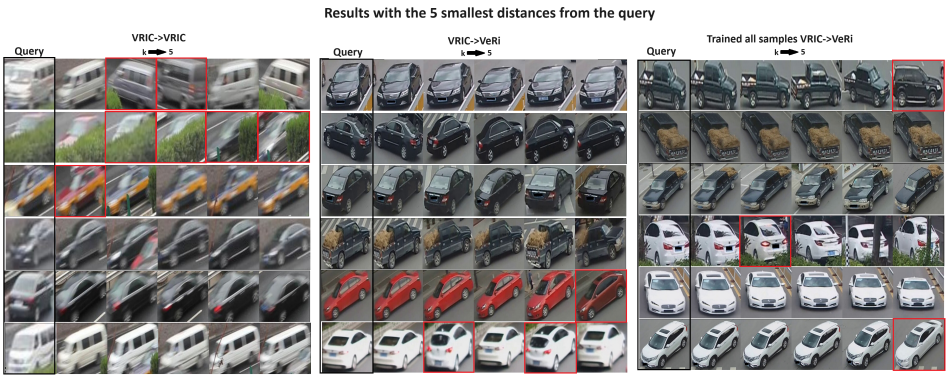

Figure 2: Qualitative inference of our model, VRIC to VeRi vs. VRIC to VRIC vs. large VRIC to VeRi.

Table 4: Experimental results of Different Models on Large dataset Re-Identification Tasks, $VRIC->VeRi$ vs $VeRi->VRIC$.

Model	CMC Top 1	CMC Top 5	CMC Top 10	Train	Test
PGAN	95.2	97.5	98.7	VeRi	VeRi
PGAN	14.3	26.1	33.0	VeRi	VRIC
PGAN	44.9	56.9	66.2	VRIC large	VeRi
PGAN	37.5	43.2	50.8	VRIC small	VeRi
DRPC (ours)	94.8	97.2	98.4	VeRi	VeRi
DRPC (ours)	39.6	53.3	60.9	VRIC small	VeRi
DRPC (ours)	**56.8**	**62.1**	**68.0**	VRIC large	VeRi
DRPC (ours)	21.8	38.2	46.4	VeRi	VRIC

perform well in similar contexts. As observed, PGAN and the proposed method are nearly identical regarding the CMC curve for VeRi vs. VeRi, with only a 0.5 percentage point difference at K=5.

Regarding susceptibility to input variability, our model's ability to learn contextual regions allowed it to achieve superior performance in the Cumulative Matching Characteristic (CMC) metric with $K = 5$ when transitioning from the VeRi dataset to the VRIC dataset (VeRi vs VRIC). Specifically, the model improved up to 12.1 percentage points, indicating enhanced generalization capabilities. In contrast, the model trained with a smaller dataset and data augmentation fell 5.3 percentage points short compared to the baseline model trained on the full dataset.

3.4 Discussion of Results

Our model focused on improving generalization and performance on small datasets. This led to evaluating the performance of our model in terms of changes in the number of samples and context variations. Figure 2 shows how the matching works for the k=5 smallest distances concerning our query, evaluated on a gallery or query set.

Figure 2 shows some critical and common cases regarding re-identification for VRIC and VeRi. Regarding the VRIC dataset, we found quite a few matches for the top 5, considering the dimensionality of the samples used in the training, which is quite positive. The shift in context to VeRi demonstrated that part of the knowledge extracted from VRIC helped to re-identify a wide variety of instances in this dataset. The scores in the context shift was deficient compared to the base model trained with attention and the entire VRIC set due to the inherent limitations of our small dataset's dimensionality. Regarding our model's effectiveness in generalizing learning in a large dataset, we also observed that our model not only reduced the necessary training time but also generalized the learning optimally, achieving a balance between performance and susceptibility to variability.

4 COPYRIGHT FORM

5 CONCLUSIONS

Our approach demonstrated how to tackle a mask-based method by creating regions through the same activations proposed by the backbone, generating partitions with greater context, and establishing a baseline for scenarios with high-resolution instances where emphasis is placed on the fusion of global and local attributes. Along the same lines, we emphasized creating more discriminative feature maps by reducing feature maps of proposed regions, leading to more representative representations per proposed region. This made our approach more robust in domain

shift and small datasets. Providing greater context and more localized information resulted in our model outperforming our baseline by 8% in the small dataset case, comparing our baseline PGAN result with our DPRC-based model in the context of VRIC small. In domain shift, our model (DPRC) exceeded our baseline (PGAN) by up to 11% in CMC k=1, showing positive effects of proposing context-rich region parts in using VRIC to train and VeRI for inference. Additionally, our approach aimed to verify the limitations involved in small sample volumes and observed how implementing classical techniques oriented towards morphological transformations and using GANs for simple changes like color and texture can help address issues where our study instance is highly costly to annotate.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordination of Superior Level Staff Improvement - Brasil (CAPES) - Finance Code 001.

REFERENCES

Chen, H., Zhao, Y., and Wang, S. (2023). Person re-identification based on contour information embedding. *Sensors*, 23(2):774.

Chen, X., Xu, H., Li, Y., and Bian, M. (2021). Person re-identification by low-dimensional features and metric learning. *Future Internet*, 13(11):289.

Fan, X., Luo, H., Zhang, X., He, L., Zhang, C., and Jiang, W. (2019). Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pages 19–34. Springer.

Fawad, Khan, M. J., and Rahman, M. (2020). Person re-identification by discriminative local features of overlapping stripes. *Symmetry*, 12(4):647.

Gong, Y., Zeng, Z., Chen, L., Luo, Y., Weng, B., and Ye, F. (2021). A person re-identification data augmentation method with adversarial defense effect. *arXiv preprint arXiv:2101.08783*.

He, X., Luo, Z., Li, Q., Chen, H., and Li, F. (2023). Dg-gan: A high quality defect image generation method for defect detection. *Sensors*, 23(13):5922.

Jiang, Y., Chen, W., Sun, X., Shi, X., Wang, F., and Li, H. (2021). Exploring the quality of gan generated images for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4146–4155.

Kalayeh, M. M., Basaran, E., Gökmen, M., Kamasak, M. E., and Shah, M. (2018). Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1062–1071.

Kanaci, A., Zhu, X., and Gong, S. (2018). Vehicle re-identification in context. In *Pattern Recognition - 40th German Conference, GCPR 2018, Stuttgart, Germany, September 10-12, 2018, Proceedings*.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*.

Liu, X., Liu, W., Mei, T., and Ma, H. (2016). A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 869–884. Springer.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., and Gu, J. (2019a). A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609.

Luo, H., Jiang, W., Zhang, X., Fan, X., Qian, J., and Zhang, C. (2019b). Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 94:53–61.

Lv, G., Ding, Y., Chen, X., and Zheng, Y. (2024). Mp2pmatch: A mask-guided part-to-part matching network based on transformer for occluded person re-identification. *Journal of Visual Communication and Image Representation*, 100:104128.

Si, J., Zhang, H., Li, C.-G., Kuen, J., Kong, X., Kot, A. C., and Wang, G. (2018). Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5363–5372.

Song, C., Huang, Y., Ouyang, W., and Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1179–1188.

Wang, G., Yuan, Y., Chen, X., Li, J., and Zhou, X. (2018). Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282.

Zhang, X., Zhang, R., Cao, J., Gong, D., You, M., and Shen, C. (2020). Part-guided attention learning for vehicle instance retrieval. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3048–3060.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., and Kautz, J. (2019). Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147.